

다국어 자연어처리 관점에서 M-BERT의 한국어 성능 평가

Evaluation of M-BERT Performance on Korean
in the Perspective of Multilingual NLP

이화여자대학교 국제학부 이송

CHAPTER 1

서론: 연구 배경, 연구 목표

연구 배경

영어에 집중되어 있는 자연어처리, 하지만 **다국어 자연어처리 모델의 필요성** 증가

-다국어 자연어처리(Multilingual language processing):
특정 언어만이 아니라 다국어 입력 데이터에 대하여 컴퓨터가 이해하고 분석할 수 있는 기술.

-다국어 자연어처리 모델의 필요 이유:

(1) **Low-resource language**도 효과적으로 학습하고 모델 만들 수 있음
→ 데이터 부족 문제 해결

(2) **Cross-lingual 테스트 수행 가능**. Cross-lingual은 서로 다른 언어에서 동일한 의미를 갖는 단어를 찾는 것.
번역 등에 주로 사용됨.
→ 예를 들면, 한-영, 한-중, 한-일 번역 모델을 각각 만들지 않고
하나의 번역모델 만으로도 가능하여
효율적이고 경제적으로 합리적임.



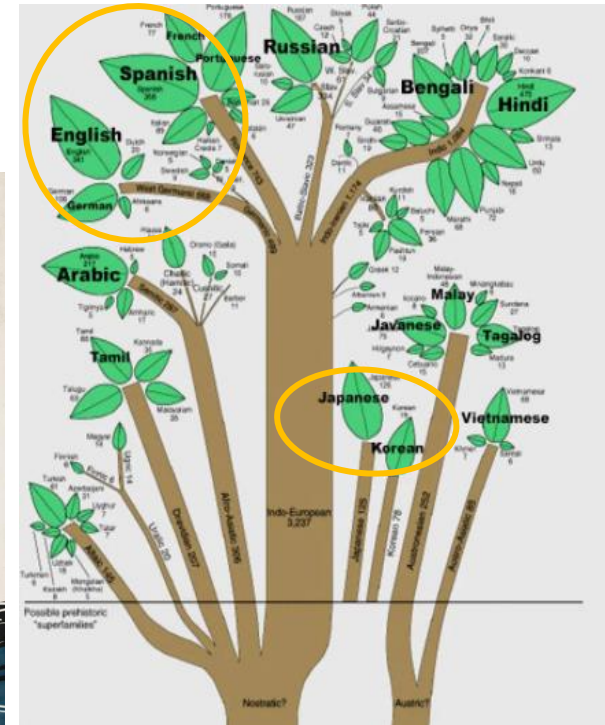
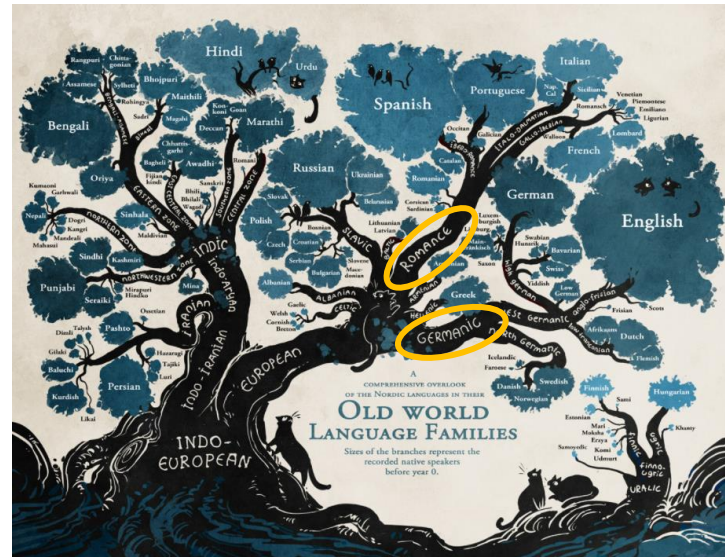
연구 배경

-대표적 다국어 자연어처리 모델: **M-BERT(BERT-Base, Multilingual Cased)**

-**Romance나 Germanic에 속하며 영어와 가까운 언어**에 대하여는 다국어 자연어처리 실험/연구가 **많이** 이루어짐
(German, French, Spanish, Italian 등)

-하지만, 유형론적(Typologically)으로 거리가 있는 **한국어나 일본어**에 대해서는 다국어 자연어처리라는 universal한 NLP 모델의 연구가 **부족한 상황**

-그리하여 다국어 자연어처리 모델에서 한국어의 성능 현황을 파악해보도록 함



Is Multilingual BERT Fluent in Language Generation?

Samuel Rönnqvist* Jenna Kanerva* Tapio Salakoski Filip Ginter*
TurkuNLP
Department of Future Technologies
University of Turku, Finland
{saanro, jmnybl, sala, figint}@utu.fi

We put a particular focus on the natural language generation (NLG) task, which we hypothesize requires a deeper understanding of the language in question on the side of the model. We take English and German, for which monolingual versions of BERT are available, as reference languages, in order to compare how they perform in the mono- vs. multilingual settings. Furthermore, we perform experiments with the Nordic languages of Danish, Finnish, Norwegian (Bokmål and Nynorsk) and Swedish, with in-depth evaluations on Finnish and Swedish, as well as the abovementioned two reference languages.

연구 목표

M-BERT에서 한국어의 성능 평가

가설1

“M-BERT에서 한국어의 성능은 유형론적으로 영어와 가까운 언어의 성능보다 낮게 나온다.”

- 각각 영어, 독일어, 한국어 데이터를 M-BERT에 넣어 성능을 비교해 봄
- 문장 내 문맥 파악 능력을 평가하는 NER과 문장 간의 문맥 파악 능력을 평가하는 QA를 수행함

가설2

“M-BERT와 KoBERT 모델을 비교했을 때, KoBERT 모델의 성능이 더 높다.”

- 범용적 다국어 자연어처리 모델인 M-BERT와 ETRI에서 한국어 데이터를 추가로 학습하여 구축한 KoBERT의 성능을 NER과 QA로 비교분석 해봄



다국어 자연어처리 모델에서 한국어의 성능을 향상시킬 수 있는 방법이 무엇이 있을지 고민해볼 수 있음.
추후 Cross-lingual approach에 있어서 많은 도움이 될 것이라 예상함.

CHAPTER 2

본론: 모델, 데이터, 실험환경 소개

딥러닝 모델: M-BERT

-BERT: 구글이 공개한 AI 언어모델. 11개 분야에서 SOTA를 달성함.

-**Bidirectional Encoder Representations from Transformer**
(트랜스포머 양방향 인코더 표현)

-**Language Representation을 해결**하기 위해 고안된 구조

Q) “단어, 문장, 언어를 어떻게 표현할까?”

A) **양방향성**을 포함하여 문맥을 더 자연스럽게 파악할 수 있다!

-Wikipedia와 BookCorpus를 학습한 후, 다양한 task에 적용하고 fine-tuning해서 사용할 수 있음

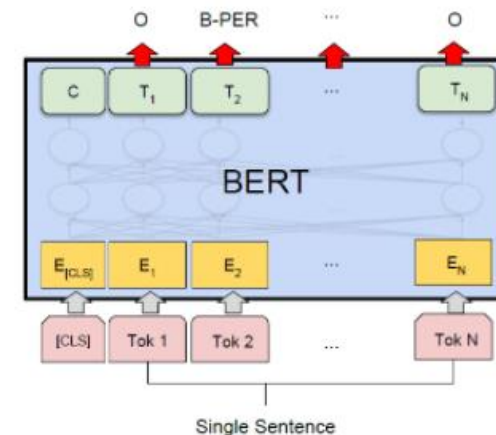
-Pre-training 시 비지도 학습 방법

1. Masked Language Model:

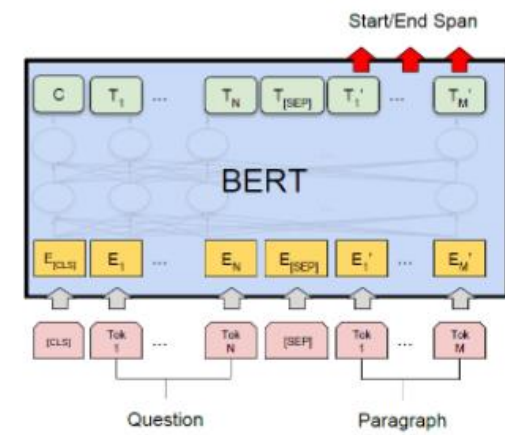
단어 중 일부를 [Mask] 토큰으로 바꾼 뒤 가려진 단어 예측하도록 학습.
문맥 파악 능력 기를

2. Next Sentence Model: 다음 문장이 올바른 문장인지 맞추는 문제.

두 문장 사이의 관계를 학습하게 됨. 문장 A와 B를 이어 붙일 때,
B는 50%확률로 관련 있는/없는 문장 사용.



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER



(c) Question Answering Tasks:
SQuAD v1.1

딥러닝 모델: M-BERT

- Multilingual BERT(M-BERT): BERT에 다국어 텍스트를 학습시킨 모델.
- BERT-Base, Multilingual Cased: Wikipedia의 104개의 언어 코퍼스로 학습됨.
- 데이터셋 분포: 영어가 21%, 독일어가 7%, 한국어가 1.1% → 불균형 존재
- Multilingual models as “Universal models”
- 존스홉킨스 대학의 논문에 따르면, M-BERT 모델은 추가적인 학습 없이도 다국어 표현을 효과적으로 학습하여 성능을 보임.
- 39의 언어에 대해 5개의 테스트를 수행함.
- 다국어 자연어처리 모델로서의 잠재력을 보여줌.

Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT

Shijie Wu and Mark Dredze
Department of Computer Science
Johns Hopkins University

shijie.wu@jhu.edu, mdredze@cs.jhu.edu

Abstract

Pretrained contextual representation models (Peters et al., 2018; Devlin et al., 2019) have pushed forward the state-of-the-art on many NLP tasks. A new release of BERT (Devlin, 2018) includes a model simultaneously pre-trained on 104 languages with impressive performance for zero-shot cross-lingual transfer on a natural language inference task. This paper explores the broader cross-lingual potential of mBERT (multilingual) as a zero-shot language transfer model on 5 NLP tasks covering a total of 39 languages from various language families: NLI, document classification, NER, POS tagging, and dependency parsing. We compare mBERT with the best-published methods for zero-shot cross-lingual transfer and find mBERT competitive on each task. Additionally, we investigate the most effective strategy for utilizing mBERT in this manner, determine to what extent mBERT generalizes away from language-specific features, and measure factors that influence cross-lingual transfer.

At the same time, cross-lingual embedding models have reduced the amount of cross-lingual supervision required to produce reasonable models; Conneau et al. (2017); Artetxe et al. (2018) use identical strings between languages as a pseudo bilingual dictionary to learn a mapping between monolingual-trained embeddings. Can jointly training contextual embedding models over multiple languages without explicit mappings produce an effective cross-lingual representation? Surprisingly, the answer is (partially) yes. BERT, a recently introduced pretrained model (Devlin et al., 2019), offers a multilingual model (mBERT) pre-trained on concatenated Wikipedia data for 104 languages without any cross-lingual alignment (Devlin, 2018). mBERT does surprisingly well compared to cross-lingual word embeddings on zero-shot cross-lingual transfer in XNLI (Conneau et al., 2018), a natural language inference dataset. **Zero-shot cross-lingual transfer**, also known as single-source transfer, refers to training and selecting a model in a source language, often a high resource language, then transfers directly to a target language.

1 Introduction

Pretrained language representations with self-supervised objectives have become standard in a variety of NLP tasks (Peters et al., 2018; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2019), including sentence-level classification (Wang et al., 2018), sequence tagging (e.g. NER) (Tjong Kim Sang and De Meulder, 2003) and SQuAD question answering (Rajpurkar et al., 2016). Self-supervised objectives include language modeling, the cloze task (Taylor, 1953) and next sentence classification. These objectives continue key ideas in word embedding objectives like CBOW and skip-gram (Mikolov et al., 2013a).

Code is available at <https://github.com/shijie-wu/crosslingual-nlp>

While XNLI results are promising, the question remains: does mBERT learn a cross-lingual space that supports zero-shot transfer? We evaluate mBERT as a zero-shot cross-lingual transfer model on five different NLP tasks: natural language inference, document classification, named entity recognition, part-of-speech tagging, and dependency parsing. We show that it achieves competitive or even state-of-the-art performance with the recommended fine-tune all parameters scheme (Devlin et al., 2019). Additionally, we explore different fine-tuning and feature extraction schemes and demonstrate that with parameter freezing, we further outperform the suggested fine-tune all approach. Furthermore, we explore the extent to which mBERT generalizes away from a specific language by measuring accuracy on language ID

833

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 833–844, Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

***** New November 23rd, 2018: Un-normalized multilingual model + Thai + Mongolian *****

We uploaded a new multilingual model which does *not* perform any normalization on the input (no lower casing, accent stripping, or Unicode normalization), and additionally includes Thai and Mongolian.

It is recommended to use this version for developing multilingual models, especially on languages with non-Latin alphabets.

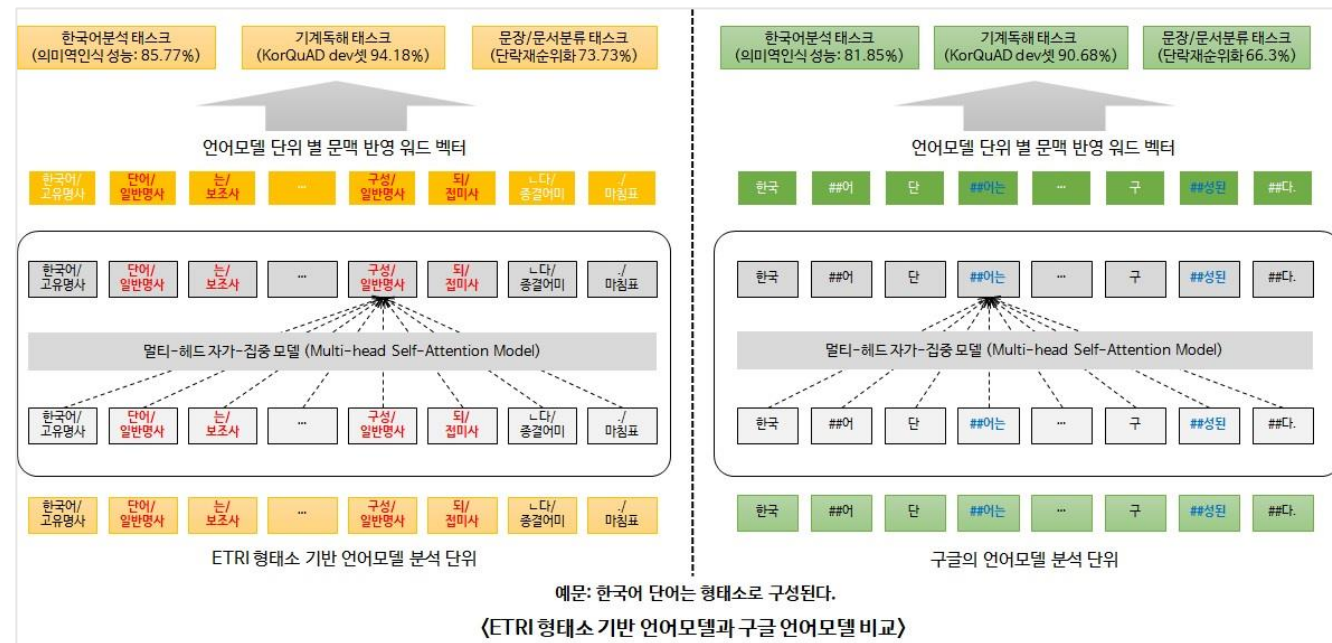
This does not require any code changes, and can be downloaded here:

- **BERT-Base, Multilingual Cased** 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters

<https://github.com/google-research/bert>

딥러닝 모델: KoBERT

- ETRI 엑소브레인 연구진이 한국어의 특성을 반영하여 개발한 BERT 언어모델
- 5종(의미역 인식, 기계독해, 단락 순위화, 문장 유사도 추론, 문서 주제분류)의 한국어 처리 태스크에서 구글이 배포한 한국어 언어모델과 비교평가 시, ETRI 모델이 평균 4.5% 성능이 우수한 것으로 평가됨
- 신문기사와 백과사전 등 **23GB의 대용량 텍스트**를 대상으로 47억개의 형태소를 사용하여 학습
- 형태소분석 기반 언어모델**과 형태소분석을 수행하지 않은 **어절 기반의 언어모델** 두 가지 제공
- Open API 형태로 되어 있어 개발급을 신청한 뒤 학습데이터를 신청해야 됨
https://aiopen.etri.re.kr/service_dataset.php



데이터셋

[NER]

Named Entity Recognition. 개체명인식.

문자열 안의 NE(명사)의 위치를 알아내고, 사전정의한 카테고리에 따라 알맞게 분류하는 것.

이름	설명	링크
CONLL2003	University of Antwerp에서 제공하는 데이터셋. 영어와 독일어 제공. 인물(PER), 기관 및 단체(ORG), 장소 및 위치(LOC)으로 이루어져 있음	https://arxiv.org/abs/cs/0306050
모두의 말뭉치 (개체명)	국립국어원에서 제공하는 데이터셋. 장소(LC), 날짜(DT), 기관(OG), 시간(TI), 인물(PS)로 이루어져 있음.	https://corpus.korean.go.kr

[QA]

Question Answering. 질의응답.

자연언어로 인간이 제기한 질문에 자동으로 응답하는 시스템 구축. 독해 능력 평가.

이름	설명	링크
SQuAD	Stanford Question Answering Dataset. 모든 질의에 대한 답변은 해당 Wikipedia 아티클 문단의 일부 하위 영역으로 이루어짐. 100K+의 데이터셋 제공	https://rajpurkar.github.io/SQuAD-explorer/
KorQuAD	LG CNS에서 100K의 데이터셋 제공	https://korquad.github.io/
GermanQuAD	독일의 Deepset에서 20K의 데이터셋 제공	https://deepset.ai/germanquad

실험환경

-언어: Python 3.9






-환경: Colab의 GPU/TPU, Tensorflow, Keras

-평가지표: 모델의 정확성 측정을 위하여 F1 Score 사용

-F1 Score: Precision과 Recall의 조화평균

Precision은 True라고 분류한 것 중 실제 True의 비율,
Recall은 실제 True인 것 중에서 모델이 True라고 예측한 비율

$$(F1-score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

multi_cased_L-12_H-768_A-12 > multi_cased_L-12_H-768_A-12			
이름	수정한 날짜	유형	크기
 bert_config.json	2021-05-06 오후 12:04	JSON File	1KB
 bert_model.ckpt.data-00000-of-00001	2021-05-06 오후 12:04	DATA-00000-OF-0...	697,526KB
 bert_model.ckpt.index	2021-05-06 오후 12:04	INDEX 파일	9KB
 bert_model.ckpt.meta	2021-05-06 오후 12:04	META 파일	888KB
 vocab.txt	2021-05-06 오후 12:04	텍스트 문서	973KB

```
[ ] !pip install keras-bert #keras-bert 케라스에서 Bert 활용을 쉽게 하는 모듈
!pip install keras-radam #Adam optimizer 수정판

[ ] #keras-bert 라이브러리에서 버트 모델 활용에 필요한 모듈 임포트
from keras_bert import load_trained_model_from_checkpoint, load_vocabulary
from keras_bert import Tokenizer
from keras_bert import AdamWarmup, calc_train_steps

from keras_radam import RAdam

[ ] #Gdrive 상에 있는 Bert 모델을 Colab 클라우드 컴퓨터 안에 저장(사전학습된 모델 로드 시간 단축 가능)
def copytree(src, dst, symlinks=False, ignore=None):
    for item in os.listdir(src):
        s = os.path.join(src, item)
        d = os.path.join(dst, item)
        if os.path.isdir(s):
            shutil.copytree(s, d, symlinks, ignore)
        else:
            shutil.copy2(s, d)

[ ] #Colab 클라우드 서버에 bert라는 폴더 생성
os.makedirs("bert")

[ ] #bert폴더에 Gdrive에서 다운받았던 bert 사전학습 모델 등의 파일을 Gdrive에서 클라우드 내부로 복사
copytree(os.path.join(path, "bert"), "bert")

[ ] #데이터를 다운로드
```

CHAPTER 3

결론 및 독창성: 연구의 의의

연구의 의의

한국어를 다국어 자연어처리 관점에서 바라보고
타 언어/모델과 비교분석함

한국에서 다국어 자연어처리 모델에 대한
연구의 활성화와 방향에 대하여 제시함